ORIGINAL

NATIONAL UNIVERSITY OF SINGAPORE

EXAMINATION FOR

(Semester 1: 2019/2020)

EE5907 – PATTERN RECOGNITION

Time Allowed: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. This paper contains **FOUR (4)** questions and comprises **FIVE (5)** printed pages.
2. All questions are compulsory. Answer **ALL** questions.
3. The total mark is **ONE HUNDRED (100)**.
4. This is a **CLOSED BOOK** examination. One A4-size formula sheet is allowed.
5. Non-programmable calculators are allowed.

Q1 (25 marks). Subquestions (a) and (b) can be answered independently.

(a) Consider a binary classification problem of predicting binary class y from features x. The cost of correct prediction is \$0. There is a \$4 cost associated with predicting class 0 when the true class is 1. There is a \$8 cost associated with predicting class 1 when the true class is 0. Suppose the cost of asking a human to perform the manual classification is \$1. Therefore, for a particular x, there are three possible decisions: (1) decision α_0 predicts y to be 0, (2) decision α_1 predicts y to be 1 and (3) decision α_h requires a human to perform the manual classification. Let $p_1 = p(y = 1|x)$

(i) Assume the human is 100% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

(ii) Assume the human is only 90% accurate. Assume that when the human is wrong, the correct class is equal to class 0 with probability 0.7 and class 1 with probability 0.3. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(7 marks)

(b) Suppose $x_n \sim \mathcal{N}(\mu, 1)$ and we observe $x_{1:N} = \{x_1, \ldots, x_N\}$. Given that $p(\mu|x_{1:N}) \sim \mathcal{N}(2, 1)$, what is the posterior predictive distribution $p(x_{N+1}|x_{1:N})$? For full credit, please show all your steps. Your final answer should not contain μ.

[Hint: $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz = 1$]

(12 marks)

Q2 (25 marks). Subquestions (a) and (b) can be answered independently.

(a) Consider a 2-class naive Bayes classifier with one binary feature and one Gaussian feature. More specifically, class label y follows a categorical distribution parametrized by π, i.e., $p(y = c) = \pi_c$. The first feature x_1 is binary and follows a Bernoulli distribution: $p(x_1|y = c) = \text{Bernoulli}(x_1|\theta_c)$. The second feature x_2 is univariate Gaussian: $p(x_2|y = c) = \mathcal{N}(x_2|\mu_c, \sigma_c^2)$. Let $\pi = [0.4\ \ 0.6]$, $\theta = [0.7\ \ 0.5]$, $\mu = [1\ \ 0]$ and $\sigma^2 = [1\ \ 1]$.

(i) Compute $p(y|x_2 = 0)$. Note that result is a vector of length 2 that sums to 1.
(7 marks)

(ii) Compute $p(y|x_1 = 0, x_2 = 0)$. Note that result is a vector of length 2 that sums to 1.
(6 marks)

(b) Suppose we toss a coin N times and observe N_1 heads. Now consider the following prior:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.7 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases}$$

Derive the MAP estimation under the prior as a function of N_1 and N.

(12 marks)

Q3 (25 marks). Suppose d-dimensional data vectors $x \in R^d$ from C classes are provided, with $n_i$vectors from class c_i for $i = 1, \ldots, C$ and $\sum_{i=1}^{C} n_i = n$.

(a) Linear discriminative analysis (LDA) projection direction, $W \in R^{d \times p}$, is obtained by maximizing

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

where $S_B \in R^{d\times d}$ and $S_w \in R^{d\times d}$ are the between class scatter and within class scatter respectively. Derive the expression for the optimal projection direction W^* that maximizes $J(W)$.

(8 marks)

(b) Marginal fisher analysis (MFA) uses neighbouring data to compute the scatter matrices S_B and S_W, and obtain the projection matrix W accordingly. Suppose k_1 neighbouring data are used for computing S_W and k_2 neighbouring data are used for computing S_B. Write the objective to optimize for MFA. Derive the optimal solution W^*.

(10 marks)

(c) Compare the MFA and LDA by listing their advantages and disadvantages. For what kind of data distribution, MFA is more preferred than LDA?

(7 marks)

Q4 (25 marks). Consider a $d - n_H - C$ fully connected neural network. The input dimension is d, number of hidden units is n_H and dimension of output is C. Answer the following sub-questions.

(a) Draw the architecture of the above network and label connection parameters between two adjacent layers on the network. Here denote the parameters as w_{ij}.

(7 marks)

(b) Non-linear activation functions are usually used between two layers. The widely used non-linear activation functions include Sigmoid and ReLU. Write their formulations and explain why ReLU is more preferred in modern neural network architectures.

(5 marks)

(c) Suppose the network is to be trained using the following loss function

$$L = \frac{1}{4}\sum_{k=1}^{n}(ReLU(t_k) - z_k)^4$$

Here z_k is a scalar and is the prediction for the input data $x_k \in R^d$ and t_k is the regression target for x_k. There are in total n data samples. Derive the learning rule Δw_{ij} for the hidden-to-output weights.

(13 marks)

END OF PAPER